Supplementary Materials

These supplementary materials present proofs of Theorems 2, 3, 4, and 5. First, we derive Theorem 2 as explained hereinafter.

Theorem 2 *Under Assumption 1, let $G^*(\mathbf{V})$ be the true structure. I-maps NPCDAG with the fewest NCP are classification equivalent to $G^*(\mathbf{V})$.*

Proof *Let $G^*_{NPC}(\mathbf{V})$ be an NCPDAG that is classification equivalent to $G^*(\mathbf{V})$. This theorem can be proven by contradiction. Assuming that there exists an I-map NPCDAG G' with the fewest NCP which is not classification equivalent to $G^*_{NPC}(\mathbf{V})$, then because G' has the fewest NCP in I-maps NPCDAG, G' represents some d-separations related with the class variable which $G^*_{NPC}(\mathbf{V})$ does not represent. Such d-separations are also not represented by $G^*(\mathbf{V})$ because $G^*(\mathbf{V})$ is classification equivalent to $G^*_{NPC}(\mathbf{V})$. This lack of representation contradicts that G' is an I-map, which completes the proof.* □

Next, we introduce the following theorem and definitions.

theorem *(Local independences in Bayesian networks) (Pearl 2000)*
Letting $G = (\mathbf{V}, \mathbf{E})$ be a Bayesian network structure, and letting $\mathbf{ND}_G(X)$ be a set of non-descendants of X, then the following holds:

$$\forall X \in \mathbf{V}, Dsep_G(X, (\mathbf{ND}_G(X) \setminus \mathbf{Pa}_X^G) \mid \mathbf{Pa}_X^G).$$

Definition *(Asymptotic consistency of scoring criterion) (Chickering 2002)*
Let $G_1 = (\mathbf{V}, \mathbf{E}_1)$, and $G_2 = (\mathbf{V}, \mathbf{E}_2)$ be the structures. A scoring criterion $Score$ has asymptotic consistency if the following two properties hold when the sample size is sufficiently large.

- *If G_1 is an I-map and G_2 is not an I-map, then $Score(G_1) > Score(G_2)$.*
- *If G_1 and G_2 both are I-maps and if G_1 has fewer parameters than G_2, then $Score(G_1) > Score(G_2)$.*

Definition *(Asymptotic local consistency of scoring criterion) (Chickering 2002)*
Let $G_1 = (\mathbf{V}, \mathbf{E}_1)$ be any structure. Also, let G_2 be the structure which results from adding edge $Y \to X$. A scoring criterion $Score$ has an asymptotic local consistency if the following two properties hold when the sample size is sufficiently large.

- $I(X, Y \mid \mathbf{Pa}_X^{G_1}) \Rightarrow Score(G_1) > Score(G_2)$.
- $\neg I(X, Y \mid \mathbf{Pa}_X^{G_1}) \Rightarrow Score(G_1) < Score(G_2)$.

To derive Theorem 3, we introduce the following lemma.

Lemma 1 *Assuming disjoint variable sets $\mathbf{X}, \mathbf{Y}, \mathbf{A}, \mathbf{B}$, then the following holds.*

$$\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \vee \neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B}).$$

Proof *From the decomposition property of conditional independence (Pearl 1988), $I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A}) \Rightarrow I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \wedge I(\mathbf{X}, \mathbf{B} \mid \mathbf{A})$ holds. The contraposition of the implication above is $\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \vee \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A})$. One obtains*

$$\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A}). \tag{1}$$

From the intersection property of conditional independence (Pearl 1988), $I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \wedge I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B}) \Rightarrow I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A})$ holds. The contraposition of the implication presented above is

$$\neg I(\mathbf{X}, (\mathbf{Y} \cup \mathbf{B}) \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \vee \neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B}). \tag{2}$$

From (1) and (2), we obtain $\neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A}) \Rightarrow \neg I(\mathbf{X}, \mathbf{B} \mid \mathbf{A} \cup \mathbf{Y}) \vee \neg I(\mathbf{X}, \mathbf{Y} \mid \mathbf{A} \cup \mathbf{B})$. □

Consequently, we derive Theorem 3 as explained below.

Theorem 3 *For a sufficiently large sample, the highest BDeu scoring structure consistent with order π is an I-map with the minimum NCP among all the structures consistent with π.*

Proof *We let $G^*_\pi = (\mathbf{V}, \mathbf{E}^*_\pi)$ be the structure with the highest BDeu among all structures consistent with order π. Also, we let $G_\pi = (\mathbf{V}, \mathbf{E}_\pi)$ be an arbitrary I-map consistent with π. From the asymptotic consistency of BDeu (Chickering 2002), G^*_π is an I-map. A sufficient condition for Theorem 3 to hold is $\mathbf{E}^*_\pi \subseteq \mathbf{E}_\pi$. This proposition can be proved as true by contradiction. Assuming that there exists an I-map consistent with π, denoted as $G'_\pi = (\mathbf{V}, \mathbf{E}'_\pi)$, such that $\mathbf{E}^*_\pi \nsubseteq \mathbf{E}'_\pi$. This assumption engenders $\exists X, Y \in \mathbf{V}, (Y \to X) \in \mathbf{E}^*_\pi \wedge (Y \to X) \notin \mathbf{E}'_\pi$. Letting $\mathbf{A} = \mathbf{Pa}_X^{G^*_\pi} \setminus \{Y\}$,* *then we obtain $\neg I(X, Y \mid \mathbf{A})$ from $(Y \rightarrow X) \in \mathbf{E}^*_\pi$ and the asymptotic local consistency of BDeu (Chickering 2002). Let $\mathbf{B}$ be a set of variables $\mathbf{Pre}^\pi_X \setminus \mathbf{Pa}^{G^*_\pi}_X$. From $\neg I(X, Y \mid \mathbf{A})$ and Lemma 1, $\neg I(X, \mathbf{B} \mid \mathbf{A} \cup \{Y\}) \vee \neg I(X, Y \mid \mathbf{A} \cup \mathbf{B})$ holds, i.e., $I(X, \mathbf{B} \mid \mathbf{A} \cup \{Y\}) \Rightarrow \neg I(X, Y \mid \mathbf{A} \cup \mathbf{B})$ holds. Because $I(X, \mathbf{B} \mid \mathbf{A} \cup \{Y\})$ holds from the local independences in G^*_π, we obtain*

$$\neg I(X, Y \mid \mathbf{A} \cup \mathbf{B}). \tag{3}$$

Also, $Dsep_{G'_\pi}(X, Y \mid \mathbf{A} \cup \mathbf{B})$ holds because X and Y are not adjacent in G'_π and because no variable in $\mathbf{A} \cup \mathbf{B}$ is a descendant of both X and Y in G'_π. This result contradicts (3), which completes the proof. □

Moreover, we derive Theorems 4 and 5 as described below.

Theorem 4 *For any variable set $\mathbf{V}$, let $G^*(\mathbf{V})$ be an I-map with minimum NCP, and let $G^{NB}(\mathbf{V})$ be the naive Bayes classifiers consisting of a set of feature variables $\mathbf{V}_c$, which are children of the class variable in $G^*(\mathbf{V})$. The following property holds.*

$$NCP(G^{NB}(\mathbf{V}_c)) \leq NCP(G^*(\mathbf{V})).$$

Proof *Because the parent of feature variables in $G^{NB}(\mathbf{V}_c)$ is only X_0, we obtain*

$$NCP(G^{NB}(\mathbf{V}_c)) = \sum_{X_i \in \mathbf{V}_c} NCP_i(\{X_0\}) + r_0 - 1,$$

where $NCP_i(\{X_0\}) = (r_i - 1)r_0$. For all $X_i \in \mathbf{V}_c$, let q^*_i be the number of parent configurations of X_i in $G^*(\mathbf{V})$. Because $X_0 \in \mathbf{Pa}^{G^*(\mathbf{V})}_{X_i}$, we obtain

$$NCP_i(\{X_0\}) \leq NCP_i(\mathbf{Pa}^{G^*(\mathbf{V})}_{X_i}).$$

Consequently, we obtain

$$\begin{aligned} NCP(G^{NB}(\mathbf{V}_c)) &= \sum_{X_i \in \mathbf{V}_c} NCP_i(\{X_0\}) + r_0 - 1 \\ &\leq \sum_{X_i \in \mathbf{V}_c} NCP_i(\mathbf{Pa}^{G^*(\mathbf{V})}_{X_i}) + r_0 - 1 \\ &= NCP(G^*(\mathbf{V})). \end{aligned}$$

□

Theorem 5 *h^* has consistency.*

Proof *For any pair of nodes $(\mathbf{U}, \mathbf{R})$ in which $\mathbf{R}$ has an incoming edge from $\mathbf{U}$ in an NROG, let $c(\mathbf{U}, \mathbf{R})$ be a cost of the edge from $\mathbf{U}$ to $\mathbf{R}$. Moreover, let X_j be a variable included in $\mathbf{U} \setminus \mathbf{R}$. When $X_j \notin \mathbf{V_c}$, we obtain*

$$\begin{aligned} h^*(\mathbf{U}) &= \sum_{X_i \in (\mathbf{U} \cup \mathbf{V}_c)} NCP_i(X_0) \\ &= \sum_{X_i \in (\mathbf{R} \cup \mathbf{V}_c)} NCP_i(X_0) \\ &\leq \sum_{X_i \in (\mathbf{R} \cup \mathbf{V}_c)} NCP_i(X_0) + NCP_j(g^*_j(\mathbf{U} \setminus \{X_j\})) \\ &= h^*(\mathbf{R}) + c(\mathbf{U}, \mathbf{R}). \end{aligned}$$

When $X_j \in \mathbf{V}_c$, the following equation holds using $X_0 \in g^*_j(\mathbf{U} \setminus \{X_j\})$.

$$\begin{aligned} h^*(\mathbf{U}) &= \sum_{X_i \in (\mathbf{U} \cup \mathbf{V}_c)} NCP_i(X_0) \\ &= \sum_{X_i \in (\mathbf{U} \cup \mathbf{V}_c) \setminus \{X_j\}} NCP_i(X_0) + NCP_j(X_0) \\ &= \sum_{X_i \in (\mathbf{R} \cup \mathbf{V}_c)} NCP_i(X_0) + NCP_j(X_0) \\ &\leq \sum_{X_i \in (\mathbf{R} \cup \mathbf{V}_c)} NCP_i(X_0) + NCP_j(g^*_j(\mathbf{U} \setminus \{X_j\})) \\ &= h^*(\mathbf{R}) + c(\mathbf{U}, \mathbf{R}). \end{aligned}$$

Consequently, we obtain

$$h^*(\mathbf{U}) \leq h^*(\mathbf{R}) + c(\mathbf{U}, \mathbf{R}).$$

□

References

Chickering, D. M. 2002. Optimal Structure Identification With Greedy Search. *JMLR*, 3: 507–554.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558604790.

Pearl, J. 2000. *Models, Reasoning, and Inference*. Cambridge University Press.